VŠB - Technická univerzita Ostrava
Fakulta elektrotechniky a informatiky
Katedra aplikované matematiky

Aproximace binomického rozdělení
Approximation of binomial distribution

2018 David Vronka

VŠB - Technická univerzita Ostrava
Fakulta elektrotechniky a informatiky
Katedra aplikované matematiky

Zadání bakalářské práce

Student: **David Vronka**

Studijní program: B2647 Informační a komunikační technologie

Studijní obor: 1103R031 Výpočetní matematika

Téma: Aproximace binomického rozdělení
Approximation of binomial distribution

Jazyk vypracování: čeština

Zásady pro vypracování:

Po podrobném popisu vlastností a možností aplikací binomického rozdělení pravděpodobnosti budou diskutovány nejčastější způsoby aproximace tohoto rozdělení. Cílem práce je také tvorba multimediálních materiálů z dané oblasti.

Seznam doporučené odborné literatury:

Grinstead ch. M.,Snell L.: Introduction to Probability, AMS, 2nd edition, 2003
Anděl J.: Matematická statistika, SNTL, 1985

Formální náležitosti a rozsah bakalářské práce stanoví pokyny pro vypracování zveřejněné na webových stránkách fakulty.

Vedoucí bakalářské práce: **Mgr. Bohumil Krajc, Ph.D.**

Datum zadání: 01.09.2017
Datum odevzdání: 30.04.2018

VYSOKÁ ŠKOLA BÁŇSKÁ-TECHNICKÁ UNIVERZITA Fakulta elektrotechniky a informatiky 1D16 OSTRAVA

doc. RNDr. Jiří Bouchala, Ph.D.
vedoucí katedry

prof. Ing. Pavel Brandštetter, CSc.
děkan fakulty

„Prohlašuji, že jsem tuto bakalářskou práci vypracoval samostatně. Uvedl jsem všechny literární prameny a publikace, ze kterých jsem čerpal.“

V Ostravě 30. dubna 2018 .....

Na tomto místě bych chtěl poděkovat panu Mgr. Bohumilu Krajcovi, Ph.D.
za pomoc při tvorbě práce.

Abstrakt

Cílem této práce je přinést nějaké znalosti o aproximaci binomického rozdělení. V první části studujeme elementární vlastnosti některých známých distribucí. Druhá část je věnována problematice odhadů minimálního počtu experimentů, založených na binomické náhodné veličině. Software byl vyvinut pro získání experimentálních výsledků z této oblasti.

Klíčová slova: binomické rozdělení, normální rozdělení, aproximace, odhad parametru, p-hodnota

Abstract

The goal of this work is to bring some knowledge about approximation of binomial distribution. In the first part we study elementary properties of some well-known distributions. The second one is devoted to the problematic of estimations of the minimal experiments based on the binomial random variable. Software was developed to obtain experimental results from this area.

Keywords: binomial distribution, normal distribution, approximation, parameter estimate, p-value

Obsah

Seznam obrázků

1. Úvod

V této bakalářské práci bylo cílem diskutovat v první řadě o užitečnosti aproximace binomického rozdělení normálním. Šlo především o souvislost se stanovením nejmenšího počtu pokusů nutných k dostatečně přesnému odhadu pravděpodobnosti úspěchu v jednom pokusu. V úvodu práce je provedeno několik intuitivních úvah o aproximaci, zejména v souvislosti s Galtonovou deskou. Práce také stručně seznamuje s některými historickými osobnostmi, které se podílely na daném problému. Teoretické základy aproximace jsou založeny na Moivreově-Laplacově větě a Bernoulliho větě, které jsou v textu formulovány. Ve druhé části se experimentálně zabýváme výpočtem nejmenšího počtu pokusů na základě různých přístupů.V práci byl použit program Matlab[1], RStudio[2] a Maple[3].

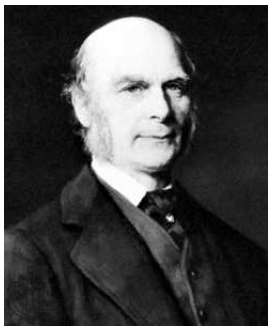
[1]MATLAB, SIMULINK, Handle Graphics a Real-Time Workshop jsou registrované známky firmy The Math Works, Inc., 3 Apple Hill Drive, Natick MA 01760-1500, USA
[2]RStudio® je registrovaná značka
[3]Maple® je registrovaná značka firmy Maplesoft

2. Historické poznámky

Studium vlastností binomického rozdělení a významné objevy v oblasti aproximace jsou spojeny se jmény tří významných matematiků. Tato část byla inspirována zejména z [WCS, WEN, WQ, JB, AM, FG].

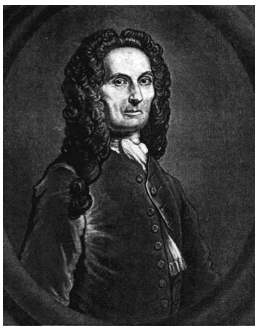
Obrázek 2.1. Jacob Bernoulli

Jacob Bernoulli byl švýcarský matematik a fyzik, narodil se 6. ledna 1655 v Basileji do rodiny belgického obchodníka Nicolause Bernoulliho. Významní členové jeho rodiny byli jeho bratr Johann Bernoulli a synovci Johann II. Bernoulli, Nicolaus II. Bernoulli a Daniel Bernoulli. Jacob přišel s myšlenkou rozdělení pravděpodobnosti, jak přesně zjistit, kolik je možných úspěchů v určitém počtu pokusů. Jde o sérii nezávislých pokusů známých jako Bernoulliho pokusy, kde může dojít pouze ke dvěma výsledkům, a to k úspěchu, nebo neúspěchu. Nejjednodušší je si to ukázat na příkladě s mincí. Zvolíme si rub jako úspěch s pravděpodobností $p = 0.5$, a tedy hod mincí může skončit úspěchem, pokud padne rub, nebo neúspěchem, padne-li líc.



Obrázek 2.2. Sir Francis Galton

Sir Francis Galton byl anglický vědec a vynálezce, který působil v různých oborech, např. v antropologii, statistice, geografii, psychologii... Narodil se 16. února 1822 v Birminghamu, byl nevlastním bratrancem Charlese Darwina. Ve statistice zavedl pojmy jako normální rozdělení, korelace a regrese. Vynalezl Galtonovu desku, o které bude pojednáno níže v části 6.1.



Obrázek 2.3. Abraham de Moivre

Abraham de Moivre byl francouzský matematik. Narodil se 26. května 1667 ve Vitry-le-François. Byl prvním, který provedl důkaz speciálního případu centrální limitní věty. Jeho klasická kniha Doktrína šancí byla v podstatě „příručka" pro gamblery, která poskytovala doporučení, jak sázet v různých hrách, kde se vyskytuje náhoda. Ve své knize Moivre například používal binomický rozvoj $(p-q)^n$ k analýze možných výsledků hodů mincí.

3. Binomické rozdělení

Řada elementárních poznatků o základních pravděpodobnostních rozděleních je obsažena např. v [AJ]. Připomeňme některé z nich.

Uvažujme experiment, jehož možným výsledkem jsou pouze dva stavy, „úspěch", nebo „neúspěch". Takovýto experiment můžeme dobře modelovat pomocí náhodné veličiny X, která se řídí tzv. alternativním rozdělením s parametrem $p \in \langle 0, 1\rangle$. V takovém případě píšeme $X \sim Alt(p)$. Formálně definujeme alternativní náhodnou veličinu X jako veličinu s oborem hodnot $\{0, 1\}$ a s vlastností:

$$Pr\,(X = 1) = p, \quad Pr\,(X = 0) = 1 - p.$$

Číslo p budeme nazývat pravděpodobností úspěchu.

Nyní uvažujme sérii n nezávislých experimentů, z nichž každý lze modelovat alternativní náhodnou veličinou se stejnou hodnotou parametru p. Takovýmto pokusům říkáme Bernoulliovy pokusy. Výsledek jednotlivého Bernoulliho pokusu je tedy pouze úspěch, nebo neúspěch. Pokud šance na úspěch je p, pak šance na neúspěch je $1 - p$. Binomické rozdělení je definováno jako náhodná veličina X, která udává počet úspěchů při n Bernoulliho pokusech. Binomické rozdělení má tedy 2 parametry. Pravděpodobnost p úspěchu v jednom pokusu a počet pokusů n. Je-li X binomická náhodná veličina s parametry n, p, pak zřejmě můžeme psát:

$$X = X_1 + X_2 + \cdots + X_n,$$

kde $X_1, X_2, \ldots, X_n$ jsou nezávislé alternativní náhodné veličiny s parametrem p.

Není těžké odvodit, že když máme n pokusů, tak pravděpodobnost, že dostaneme přesně k úspěchů, je dána vztahem:

$$Pr\,(X = x) = \binom{n}{k} p^k\,(1 - p)^{n-k}\,. \tag{3.1}$$

Většina pravděpodobnostních rozdělení má dvě nejdůležitější charakteristiky, a tím jsou střední hodnota a rozptyl. Pro zajímavost ještě uvedeme další dvě známé charakteristiky, šikmost a špičatost.

Střední hodnota binomického rozdělení:

$$E(X) = np.$$

Rozptyl:

$$D(X) = np\,(1-p)\,.$$

.

Koeficient šikmosti:

$$\gamma_1 = \frac{1-2p}{\sqrt{np\,(1-p)}}.$$

.

Koeficient špičatosti:

$$\gamma_2 = \frac{1-6p\,(1-p)}{np\,(1-p)}.$$

Řídí-li se náhodná veličina X binomickým rozdělením s parametry n a p, budeme psát $X \sim Bin\,(n,p)$.

Příklad:

Mějme 100 nakažených prasat, do kterých byly vpíchnuty injekce s protilátkou. Každé prase má 35% šanci na to, že protilátka zabere. Jaká je pravděpodobnost, že vyléčených prasat bude méně, nebo rovno 40?

Řešení:

$\sum_{i=0}^{40} Pr(X=i) \implies \sum_{i=0}^{40} \binom{100}{i} \cdot 0.35^i \cdot 0.65^{100-i} \doteq 0.8749$**.**

Pravděpodobnost toho, že vyléčených prasat nebude více než 41, je přibližně 0.875.

4. Normální rozdělení

Normální rozdělení nebo také Gaussovo rozdělení, podle Sira Friedricha Gausse, který toto rozdělení objevil, je jedno z nejdůležitějších rozdělení spojité náhodné veličiny.

Řekneme, že se náhodná veličina X řídí normálním rozdělením s parametry μ a σ , jestliže hustota pravděpodobnosti náhodné veličiny X je dána vzorcem

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad x \in \mathbb{R}.$$

Střední hodnota normálního rozdělení:

$$E(X) = \mu.$$

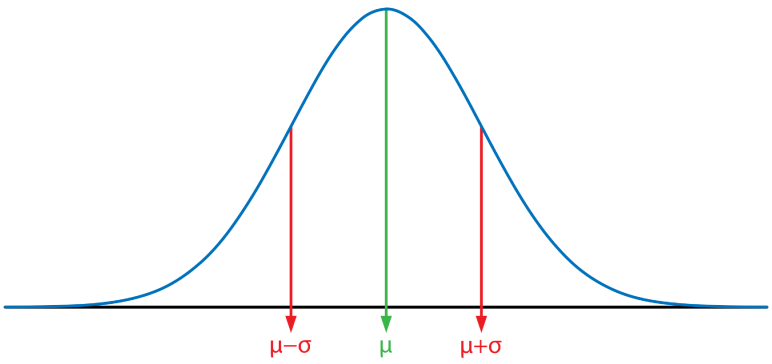
Rozptyl:

$$D(X) = \sigma^2.$$

Šikmost a špičatost:

$$\gamma_1 = \gamma_2 = 0.$$

Řídí-li se náhodná veličina X normálním rozdělením s parametry μ a σ, budeme psát $X \sim N(\mu, \sigma)$.



Obrázek 4.1. Ilustrace hustoty normálního rozdělení

Připomeňme známá fakta o normálním rozdělení (viz např. [ML]).

Grafem hustoty pravděpodobnosti náhodné veličiny s normálním rozdělením je tzv. Gaussova křivka (Gaussův klobouk,

zvonová funkce, angl. „bell curve", obrázek 4.1). Jde o zvonovitou funkci dosahující maxima pro $x = \mu$. Parametr σ odpovídá „horizontální" vzdáleností inflexních bodů od μ a tím i šířce Gaussovy křivky.

Normálním rozdělením pravděpodobnosti se například často řídí chyba měření. Má-li měřená veličina hodnotu μ, pak výsledky měření nabývají náhodných hodnot, které se řídí normálním rozdělením pravděpodobnosti, přičemž hodnota parametru σ odpovídá přesnosti přístroje.

Nechť je dána náhodná veličina $X \sim N(0,1)$. Pak pravděpodobnost, že hodnota X padne do intervalu (a,b), je dána vztahem:

$$Pr(X \in (a,b)) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_a^b \mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}} \mathrm{d}x.$$

V řadě programovacích jazyků jsou implementovány algoritmy, které umožňují přibližně vypočítat hodnoty tzv. distribuční funkce F náhodné veličiny, která se řídí normálním rozdělením s parametry μ a σ:

$$F(x) = Pr(X \in (-\infty, x)) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^x \mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}} \mathrm{d}x.$$

Příklad:

Vyjádřeme pravděpodobnost, že $X \sim N(2,4)$ nabývá hodnoty z intervalu $\langle -1, 5\rangle$ pomocí tzv. standardizované normální náhodné veličiny $Y \sim N(0,1)$.

Řešení:

$Y = \frac{X-\mu}{\sqrt{\sigma^2}} = \frac{X-2}{2} \ldots Y \sim N(0,1)$,

$X \in \langle -1, 5\rangle \Leftrightarrow Y \in \left\langle -\frac{3}{2}, \frac{3}{2}\right\rangle$,

$Pr(X \in \langle -1,5\rangle) = \frac{1}{\sqrt{2\pi 2^2}} \int_{-1}^5 \mathrm{e}^{-\frac{(x-2)^2}{2\cdot 2^2}} \mathrm{d}x = \frac{1}{\sqrt{2\pi}} \int_{-\frac{3}{2}}^{\frac{3}{2}} \mathrm{e}^{-\frac{1}{2}t^2} \mathrm{d}t \doteq 0.8664.$

Pravděpodobnost toho, že náhodná veličina $X \sim N(2,4)$ nabývá hodnot z intervalu $\langle -1, 5\rangle$, je přibližně 0.866.

5. Poissonovo rozdělení

Je rozdělení pravděpodobnosti s parametrem $\lambda t > 0$, kde $t \in \mathbb{R}^+$, $\lambda \in \mathbb{N} \cup \{0\}$. Řídí-li se náhodná veličina X Poissonovým rozdělením s parametrem λt, píšeme $X \sim Po(\lambda t)$. Hodnoty $X \sim Po(\lambda t)$ odpovídají možným výsledkům tzv. Poissonova procesu. V podstatě jde o počet událostí vykonaných v určitém časovém intervale $(0, t)$, přičemž intenzita výskytu těchto událostí je konstantní. Poissonovo rozdělení pravděpodobnosti je dáno vztahem:

$$Pr(X = k) = \frac{(\lambda t)^k}{k!} e^{-\lambda t}.$$

Parametr λ nazýváme intenzitou.

Střední hodnota:

$$E(X) = \lambda t.$$

Rozptyl:

$$D(X) = \lambda t.$$

Koeficient šikmosti:

$$\gamma_1 = \frac{1}{\sqrt{\lambda t}}.$$

Koeficient špičatosti:

$$\gamma_2 = \frac{1}{\lambda t}.$$

Příklad:

Předpokládejme, že firma Hryundai vyrobí v průměru za hodinu 60 aut. Zjistěte, jaká je pravděpodobnost, že počet vyrobených aut bude větší než 65.

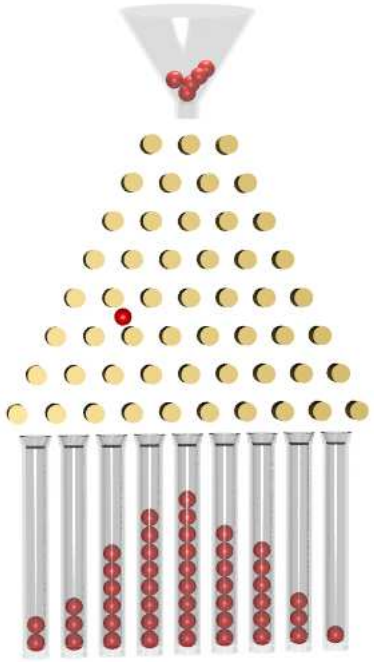
Řešení:

$Pr(X > 65) \implies 1 - \sum_{k=0}^{65} \frac{60^k}{k!} e^{-60} \doteq 1 - 0.76449351 \doteq 0.235506.$

Pravděpodobnost toho, že vyrobených aut bude více než 65, je přibližně 0.236.

Zde nejprve začněme intuitivním pozorováním.

6.1. **Galtonova deska.** Galtonova deska, též známá jako fazolový stroj, byla vynalezena Sirem Francisem Galtonem, který se snažil demonstrovat centrální limitní větu, zvláště skutečnost, že normální rozdělení je aproximací rozdělení binomického.



Obrázek 6.1. Galtonova deska

Na obrázku 6.1 (zdroj [GD]) vidíme, že kulička padá shora a má možnost spadnout doprava, nebo doleva se stejnou pravděpodobností $p = 0.5$. Po dostatečném počtu hodů kuliček si můžeme všimnout, že struktura uložených kuliček vypadá jako Gaussova křivka (viz obrázek 4.1).

6.2. **Program 1.** Pro tuto bakalářskou práci byl vytvořen program v Maplu, který vygeneruje výsledek vhazování n kuliček do Galtonovy desky 6.1 s r řádky.

Nechť vektor A představuje rozmístění kuliček (četností) na základně Galtonovy desky po provedení experimentu. Vedlejší procedura *preved* tomuto

vstupnímu vektoru A přiřadí vektor B, který převede A na odpovídající soustavu virtuálních realizací jednotlivých experimentů s kuličkami.

```
preved := proc (A)
local i, B, n, k, spocti, l;
#procedura, ktera prijima vector A;
n := Size(A, 1);
#lokalni promenna n je velikost vektoru A;
spocti := 0;
#lokalni promenna spocti, pocatecni hodnota 0;
for l to n do
spocti := spocti+A[l];
end do;
#iteracni pocitani slozek vektoru A;
B := Vector(spocti);
#lokalni promenna B, vektor o velikosti spocti;
k := 1;
#lokalni promenna k, pocatecni hodnota 1;
for i to n do
while 1 <= A[i] do
B[k] := i;
k := k+1;
A[i] := A[i]-1;
end do;
end do;
#cykly, ktere prevadi vektor A na vektor B,
#napr[1,3,2]->[1,2,2,2,3,3];
return B;
#výstup procedury
end proc
```

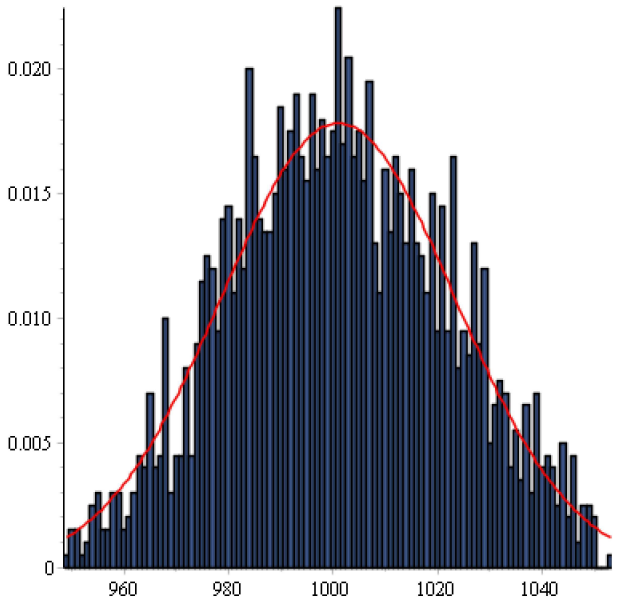
Následuje hlavní procedura *deska*, která vygeneruje graf se simulací pokusů s Galtonovou deskou, graf hustoty normálního rozdělení a otestuje normalitu dat pomocí Shapiro-Wilkova testu normality.

```
deska := proc (n, r)
local i, k, roll, A, j, g, p, P, Q, L;
#procedura deska prijima 2 parametry:
#n - pocet „vhozenych kulicek”,
```

```
#r - pocet „radku'' Galtonovy desky;
roll := rand(0 .. 1);
#lokalni promenna roll, ktera generuje cislo 0 a 1;
A := Vector(r+1);
# A vytvori nulovy vektor o velikosti r+1;
for i to n do
k := 1;
for j to r do
if irem(roll(), 2) = 0 then
k := k+1;
else
next;
end if;
end do;
#vnitrni cyklus „1 projeti kulicky''
#do posledniho radku Galtonovy desky;
A[k] := A[k]+1;
#na danou pozici v radku pricteme 1;
end do;
#''hodime n-krat kulicku'' a nastavujeme promennou k na 1;
g := preved(A);
#pouzijeme proceduru preved
#a ulozime do lokalni promenne g;
p := [seq(.5 .. r+1.5, 1)];
#vytvori seznam(vektor) cisel od 0.5
#do r+1.5 se skokem 1;
Q := Histogram(g, binbounds = p);
#funkce Histogram z knihovny Statistics;
P := DensityPlot(Normal((1/2)*r+1,
sqrt((1/4)*r+1/2)), color = "Red");
#lokalni promenna P,
#ktera do sebe ulozi hustotu normalniho rozdeleni;
print(plots[display](P, Q));
#vykresli P a Q do jednoho grafu;
L := ShapiroWilkWTest(g, level = 0.5e-1);
return L;
end proc
```

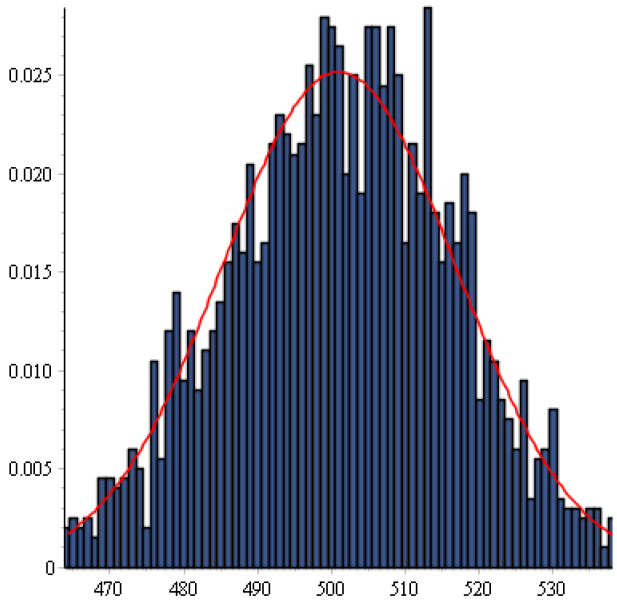
Po spuštění programu získáme například následující výstup.

```
> s := deska(2000, 2000);
Histogram Type:  variable width
Data range:      .5 .. 2001.5
Number of Bins:  2001
Frequency Scale: relative
Shapiro and Wilk's W-Test for Normality
---------------------------------------
Null Hypothesis:
Sample drawn from a population that follows
a normal distribution
Alt. Hypothesis:
Sample drawn from population that does not follow
a normal distribution
Sample size:              2000
Computed statistic:       0.98703
Computed pvalue:          0.186742
Result: [Accepted]
There is no statistical evidence against the null hypothesis
```



Obrázek 6.2. Výstup 1

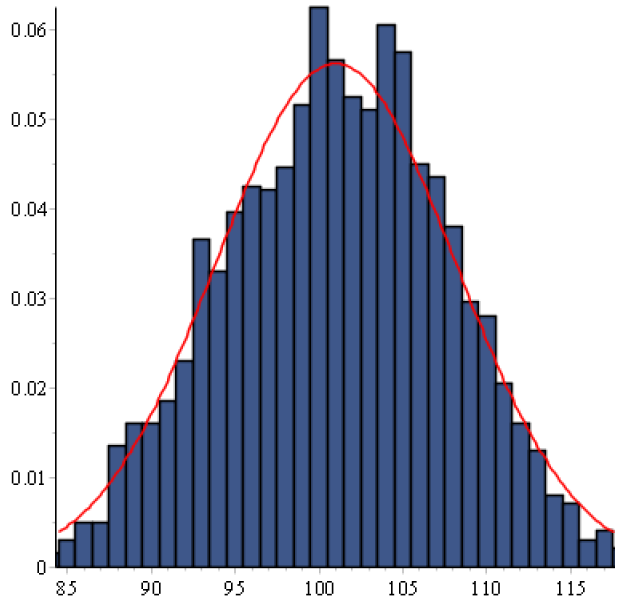
```
> s := deska(2000, 1000);
Histogram Type:  variable width
Data range:      .5 .. 1001.5
Number of Bins:  1001
Frequency Scale: relative
Shapiro and Wilk's W-Test for Normality
---------------------------------------
Null Hypothesis:
Sample drawn from a population that follows
a normal distribution
Alt. Hypothesis:
Sample drawn from population that does not follow
a normal distribution
Sample size:              2000
Computed statistic:       0.98727
Computed pvalue:          0.236718
Result: [Accepted]
There is no statistical evidence against the null hypothesis
```



Obrázek 6.3. Výstup 2

```
Histogram Type:  variable width
```

```
Data range:       .5 .. 201.5
Number of Bins:   201
Frequency Scale: relative
Shapiro and Wilk's W-Test for Normality
---------------------------------------
Null Hypothesis:
Sample drawn from a population that follows
a normal distribution
Alt. Hypothesis:
Sample drawn from population that does not follow
a normal distribution
Sample size:              2000
Computed statistic:       0.988032
Computed pvalue:          0.428579
Result: [Accepted]
There is no statistical evidence against the null hypothesis
```



Obrázek 6.4. Výstup 3

U předešlých obrázků, když jsme testovali normalitu dat, jsme uvažovali nulovou hypotézu, že data pocházejí z normálního rozdělení na hladině významnosti $\alpha = 0.05$. Připomeňme, že pokud vyjde p-hodnota větší nebo rovna než

hladina významnosti α , pak nezamítáme nulovou hypotézu.V našem případě nezamítáme nulovou hypotézu, jelikož nemáme dostatečné informace o tom, abychom ji mohli zamítnout.

V následujícím odstavci můžeme vidět Shapiro-Wilkův test provedený nad stejnými daty, jako jsou na předešlých obrázcích, ale v jiném programu nazývaném RStudio se stejnou hladinou významnosti $\alpha = 0.05$. Avšak nyní dostáváme jiné výsledné hodnoty p-hodnot. U třetího testu dokonce zamítáme nulovou hypotézu, že data pocházejí z normálního rozdělení.

```
> shapiro.test(data1$V1)
Shapiro-Wilk normality test
data:  data1$V1
W = 0.99874, p-value = 0.1535
> shapiro.test(data2$V1)
Shapiro-Wilk normality test
data:  data2$V1
W = 0.99874, p-value = 0.1555
> shapiro.test(data3$V1)
Shapiro-Wilk normality test
data:  data3$V1
W = 0.99704, p-value = 0.0007018
```

data	Maple	RStudio
1.	0.1867	0.1535
2.	0.2367	0.1555
3.	0.4286	0.0007

TABULKA 1. porovnávání hodnot p-value

Podivné výsledky vykazoval Maple i v řadě jiných situací, tykajících se statistického balíčku.

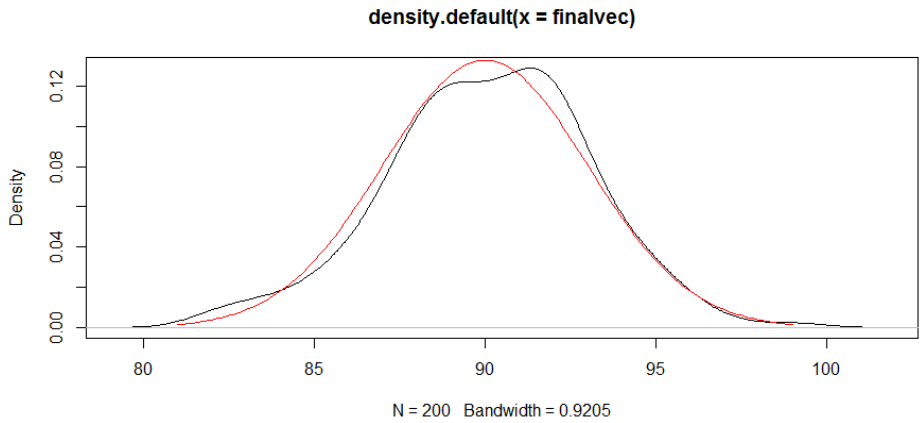
6.3. **Program 2.** Druhým programem pro tuto bakalářskou práci byla vytvořena simulace modifikované Galtonovy desky v programu RStudio.

```
GaltonBoard<-function(beans,lvls,prob)
{
  #modifikovana Galtonova deska,
  #ktera prijima 3 parametry
  #beans = pocet kulicek,
```

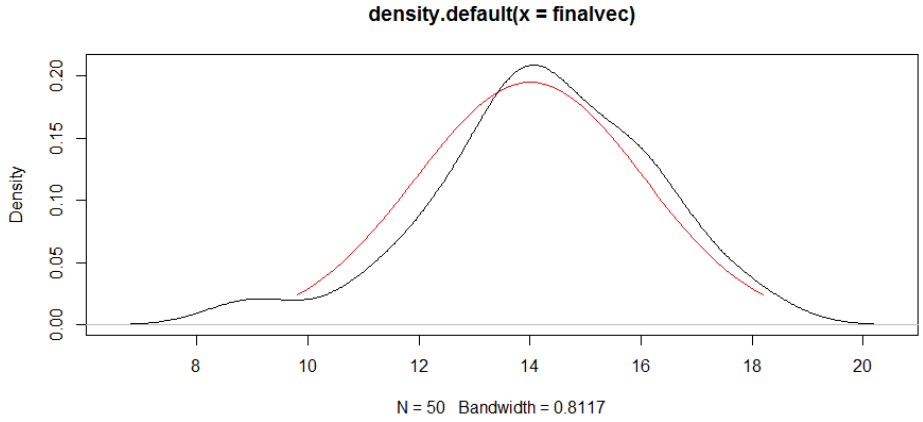
```
  #lvls = pocet radku,
  #prob = pravdepodobnost, ze kulicka pujde doleva
  finalvec<-vector()
  #vytvori prazdny vector
  for(i in 1:beans)
  {
    k<-0
    #nastaví hodnotu k na 0
    for(j in 1:lvls)
    {
      run=runif(1,0,1);
      # funkce, ktera mi vygeneruje pravdepodobnost
      #v rozsahu<0,1>
       if(run<=prob)
      {
        k=k+1;
      }
      #pokud vygenerovana hodnota bude
      #mensi/rovna nez vstupni,
      #pak spadne doprava
    }
    #cyklus pro "spadnuti" 1 kulicky
    finalvec<-c(finalvec,k)
    #prirazovani do finalvec cislo k
  }
  #cyklus pro "vhazovani" kulicek
  finalvec=sort(finalvec,decreasing=FALSE)
  #vzestupni usporadani vektoru
  x<-seq(-1,1,length=beans)*(prob*lvls*(1-prob))+lvls*prob;
  hx<-dnorm(x,prob*lvls,sqrt(prob*(1-prob)*lvls));
  plot(density(finalvec));
  lines(x,hx,col="red");
  #generovani grafu
}
```

Následují dva obrázky vygenerované modifikovanou simulací Galtonovy desky. Obrázek 6.5 byl vygenerován s parametry $beans = 200, lvls = 100, prob = 0.9$ a obrázek 6.6 se vygeneroval při zadání $beans = 50, lvls = 20, prob = 0.7$.

Červenou barvou je znázorněna hustota odpovídajícího normálního rozdělení.



Obrázek 6.5. Výstup 1 modifikované desky



Obrázek 6.6. Výstup 2 modifikované desky

Z obrázků je patrné, že pro dosti velká n a hodnoty p z rozumného rozsahu je hustota binomického rozdělení dobře aproximována hustotou normálního rozdělení. Uvedené zkoumání je však nutné považovat za intuitivní. V následující části bude zformulován klasický výsledek o aproximaci binomické rozdělení rozdělení normálním.

7. Moivreova-Laplaceova věta

V teorii pravděpodobnosti, Moivreova-Laplaceova věta, která je speciálním případem centrální limitní věty, tvrdí to, že normální rozdělení může být použito jako aproximace binomického rozdělení při splnění určitých podmínek. Zejména věta 1 ukazuje, že distribuční funkci náhodné veličiny $X_n \sim Bin(n,p)$ lze pro dosti velká n dobře aproximovat distribuční funkcí normálního rozdělení se střední hodnotou np a rozptylem $\sqrt{np(1-p)}$. V praxi se často pracuje s podmínkou:

$$np > 5 \wedge n(1-p) > 5,$$

nebo

$$np(1-p) > 9.$$

Věta 1. *(Moivreova-Laplaceova) Nechť pro každé $n \in \mathbb{N}$ je dána náhodná veličina $X_n \sim Bin(n,p)$. Utvořme posloupnost normovaných náhodných veličin Y_n,*

$$Y_n = \frac{X_n - E(X_n)}{\sqrt{D(X_n)}} = \frac{X_n - np}{\sqrt{np(1-p)}}.$$

Pak pro každé $y \in \mathbb{R}$ platí vztah

$$\lim_{n\to\infty} P(Y < y) = \Phi(y),$$

kde $\Phi(y)$ je distribuční funkce normovaného normálního rozdělení $N(0,1)$.

Jak uvedená věta plyne z centrální limitní věty, je například uvedeno v [ZR] na str. 99.

Příklad: Nechť $X \sim Bin(80, 0.2)$. Potom $E(X) = 16$ $D(X) = 12.8$ $\sqrt{D(X)} \doteq 3.58$.

$$Pr(X \in \langle 30, 50\rangle) = \sum_{k=30}^{50} \binom{80}{k} \cdot 0.2^k \cdot 0.8^{80-k} \doteq \frac{1}{\sqrt{2\pi}} \int_{3.91}^{9.5} \mathrm{e}^{-\frac{1}{2}t^2} \mathrm{d}t \doteq 4.6 \cdot 10^{-5}.$$

8. Aproximace binomického rozdělení Poissonovým

Pro ty situace, ve kterých je n velké ($n > 30$) a p je velmi malé, lze binomické rozdělení aproximovat Poissonovým. Čím je větší n a menší p, tím je aproximace lepší.

Připomeňme si střední hodnotu a rozptyl Poissonova a binomického rozdělení v tabulce 2.

Rozdělení	$E(X)$	$D(X)$
Poissonovo	λ	λ
Binomické	np	$np(1-p)$

Tabulka 2. Připomenutí

Z předchozí tabulky vidíme, že Poissonovo rozdělení má stejnou střední hodnotu a rozptyl, ale u binomického rozdělení to tak není, uvažujme proto konkrétní případ, kdy $p = 0.001$ a $n = 1000$, pak $E(X) = 1$ a $D(X) = 0.999$, což je relativně malý rozdíl, a proto budeme nahrazovat $\lambda = np$. Potom Poissonovo rozdělení pravděpodobnosti budeme brát jako:

$$Pr(x) \doteq \frac{\mathrm{e}^{-np}(np)^x}{x!}.$$

Nyní porovnejme výsledky, když použijeme binomické rozdělení a jeho Poissonovu aproximaci s parametry $n = 10000$ a $p = 0.005$. Stanovme $x = 55$.

Binomické rozdělení:

$\sum_{i=0}^{55} Pr(X = i) = \sum_{i=0}^{55} \binom{10000}{i} 0.005^i (1 - 0.005)^{10000-i} = 0.785.$

Poissonovo rozdělení:

$\sum_{i=0}^{55} Pr(X = i) = \sum_{i=0}^{55} \frac{\mathrm{e}^{10000 \cdot 0.005} \cdot (10000 \cdot 0.005)^i}{i!} = 0.7845.$

Rozdíl těchto výsledků je poměrně malý: 0.06%.

9. Bernoulliova věta

Pomocí níže uvedené věty lze zdůvodnit bodový odhad pravděpodobnosti p binomické náhodné veličiny $X_n \sim Bin\,(n,p)$ pomocí $\frac{X_n}{n}$. Nejprve si ovšem zavedeme pojem konvergence podle pravděpodobnosti.

Definice 2. Nechť je dána posloupnost náhodných veličin $X_1, X_2, \ldots, X_n, \ldots$. Řekneme, že posloupnost $\{X_n\}$ konverguje podle pravděpodobnosti k hodnotě $x \in \mathbb{R}$, (značíme $X_n \xrightarrow{\mathcal{P}} x$), jestliže pro každé $\varepsilon > 0$ platí:

$$\lim_{n\to\infty} Pr\left(|X_n - x| < \varepsilon\right) = 1.$$

Věta 3. *(Bernoulliova) Nechť je dána posloupnost náhodných veličin $X_1, X_2, \ldots, X_n, \ldots$, Nechť pro každé $n \in \mathbb{N}$ platí:*

$$X_n \sim Bin(n,p).$$

Potom platí[4]*:*

$$\frac{X_n}{n} \xrightarrow{\mathcal{P}} p.$$

Poznámka 4. Dalším důvodem pro použití aproximace parametru p číslem $\frac{X_n}{n}$ je, že jde o maximálně věrohodný odhad p.

V další části textu se budeme zabývat problematikou úzce spjatou s intervalovými odhady p. Budeme hledat nejmenší $n \in \mathbb{N}$ tak, abychom lokalizovali hodnotu p se zadanou přesností a s danou spolehlivostí.

[4] viz. např. [GS] str.308

10. Odhady minimálního počtu Bernoulliových pokusů

V této části práce naši pozornost budeme věnovat klasické, ale stále aktuální, problematice odhadu nejmenšího počtu n experimentů (počtu Bernoulliho pokusů) nutných k odhadu p_n parametru p se zadanou přesností. Ukazuje se totiž, že v řadě situací, ve kterých je cena experimentu vysoká (např. destrukční zkoušky, zdravotnický výzkum), je určení minimálního počtu experimentů naprosto zásadní. Nechť X_n udává počet úspěchů v n Bernoulliho pokusech při (neznámé) pravděpodobnosti úspěchu p v jednom pokusu. V dalším textu budeme pracovat výlučně s odhadem $p_n = \frac{X_n}{n}$ parametru p zmíněného binomického rozdělení. Tato volba je přirozená, neboť p_n představuje nejvíce věrohodný odhad p. Konkrétně, nechť jsou dána čísla $\varepsilon, \beta, \gamma$ taková, že pro odhad p_n parametru p binomického rozdělení platí:

$$p_n - \alpha < p < p_n + \beta, \text{ tj.} \tag{10.1}$$

$$-\alpha < p - p_n < \beta,$$

kde $\alpha + \beta < 2\varepsilon$. Pak řekneme, že $(p_n - a, p_n + b)$ je ε-ovým odhadem p. O symetrickém ε-ovém odhadu mluvíme v případě odhadu tvaru:

$$p_n - \varepsilon < p < p_n + \varepsilon, \text{ tj.}$$

$$|p_n - p| < \varepsilon. \tag{10.2}$$

Je-li dána hladina významnosti hodnotou $\alpha \in (0, 1)$, pak řekneme, že ε-ový odhad parametru p binomického rozdělení je dán se spolehlivostí $1-\alpha$, pokud platí, že (10.1) nastane s pravděpodobností alespoň $1 - \alpha$, tj.

$$Pr\left(-\alpha < p - p_n < \beta\right) > 1 - \alpha,$$

respektive, pokud pro symetrický odhad (10.2) platí

$$Pr\left(|p_n - p| < \varepsilon\right) > 1 - \alpha.$$

Budiž dána hladina významnosti $\alpha \in (0, 1)$. Hlavním cílem je nalézt nejmenší přirozené číslo n takové, aby příslušný intervalový odhad p byl dán alespoň se spolehlivostí $1 - \alpha$. Z univerzální Čebyševovy nerovnosti lze pro případ

symetrického odhadu (10.2) získat tzv. Bernoulliovu nerovnost pro nejmenší n ve tvaru:[5]

$$n > \frac{1}{4\varepsilon^2}.$$

Uvedený odhad lze sice vylepšit Černovovou nerovností[6]:

$$n > \frac{\ln \frac{2}{\alpha}}{2\varepsilon^2},$$

přesto jsou však uvedené odhady příliš konzervativní.

Velmi široce používanou metodou odhadu nejmenšího n je metoda založená na aproximaci binomického rozdělení pomocí normálního rozdělení. Nechť je dána náhodná veličina $X \sim Bin(n, p)$. Z Moivreovy-Laplaceovy věty plyne, že distribuční funkce náhodné veličiny

$$Y = \frac{X - np}{\sqrt{np\,(1-p)}}$$

může být aproximována distribuční funkcí ϕ rozdělení $N\,(0, 1)$. Z toho vyplývá, že pro zadanou přesnost $\varepsilon \in (0, 1)$ a hladinu významnosti α postupně platí:

$$Pr\left(\left|\frac{X}{n} - p\right| < \varepsilon\right) > 1 - \alpha \iff$$
$$Pr\left(|X - np| < n\varepsilon\right) > 1 - \alpha \iff$$

$$Pr\left(\left|\frac{X - np}{\sqrt{np\,(1-p)}}\right| < \frac{n\varepsilon}{\sqrt{np\,(1-p)}}\right) > 1 - \alpha \iff$$

$$Pr\left(|Y| < \frac{n\varepsilon}{\sqrt{np\,(1-p)}}\right) > 1 - \alpha.$$

Nyní využijeme skutečnosti, že distribuční funkce náhodné veličiny Y může být aproximována distribuční funkcí normalizovaného normálního rozdělení:

$$\Phi\,(y) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{y} e^{-\frac{-x^2}{2}} \mathrm{d}x.$$

[5]Viz například str. 27 v [RJ].
[6]Viz například zmínka v [XC] na str. 3

$$1-\alpha < Pr\left(|Y| < \frac{n\varepsilon}{\sqrt{np\,(1-p)}}\right) \approx \frac{1}{\sqrt{2\pi}} \int_{-\frac{n\varepsilon}{\sqrt{np(1-p)}}}^{\frac{n\varepsilon}{\sqrt{np(1-p)}}} e^{-\frac{-x^2}{2}}\,\mathrm{d}x =$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\frac{n\varepsilon}{\sqrt{np(1-p)}}}^{\frac{n\varepsilon}{\sqrt{np(1-p)}}} e^{-\frac{-x^2}{2}}\,\mathrm{d}x = 1 - \frac{2}{\sqrt{2\pi}} \int_{\frac{n\varepsilon}{\sqrt{np(1-p)}}}^{\infty} e^{-\frac{-x^2}{2}}\,\mathrm{d}x =$$

$$= 1 - \frac{2}{\sqrt{2\pi}} \int_{\frac{n\varepsilon}{\sqrt{np(1-p)}}}^{\infty} e^{-\frac{-x^2}{2}}\,\mathrm{d}x = 1 - 2\left[1 - \Phi\left(\frac{n\varepsilon}{\sqrt{np\,(1-p)}}\right)\right],$$

Porovnáním počáteční levé strany s konečnou pravou stranou získáváme vztah:

$$\frac{\alpha}{2} > 1 - \Phi\left(\frac{n\varepsilon}{\sqrt{np\,(1-p)}}\right).$$

Odtud plyne:

$$\Phi\left(\frac{n\varepsilon}{\sqrt{np\,(1-p)}}\right) > 1 - \frac{\alpha}{2} \text{ tzn.}$$

$$\frac{n\varepsilon}{\sqrt{np\,(1-p)}} > Z_{1-\frac{\alpha}{2}},$$

kde Z_p označuje hodnotu p kvantilu normalizovaného normálního rozdělení. Z toho po elementárních výpočtech postupně získáváme:

$$\frac{n^2\varepsilon^2}{np\,(1-p)} > Z^2_{1-\frac{\alpha}{2}}, \qquad n > \frac{p\,(1-p)\,Z^2_{1-\frac{\alpha}{2}}}{\varepsilon^2}.$$

Označíme-li, pro reálné číslo x, symbolem $\lceil x \rceil$ nejmenší celé číslo, které je větší nebo rovno než x, můžeme již z předchozích úvah snadno odvodit, že pro

$$n = \left\lceil \frac{p\,(1-p)\,Z^2_{1-\frac{\alpha}{2}}}{\varepsilon^2} \right\rceil \tag{10.3}$$

bude platit:

$$Pr\left(\left|\frac{X}{n} - p\right| < \varepsilon\right) > 1 - \alpha.$$

10.1. **Problematika přímých odhadů.** Pokud se pokusíme o přímý ε-ový odhad parametru p pomocí $\frac{X}{n}$, získáváme vztah:

$$Pr\left(\left|\frac{X}{n} - p\right| < \varepsilon\right) = Pr\left(|x - np| < n\varepsilon\right) = \sum_{x=x_1}^{x_2} \binom{n}{x} p^x\,(1-p)^{n-x},$$

kde

$$\left|\frac{x_1}{n} - p\right| < \varepsilon, \left|\frac{x_2}{n} - p\right| < \varepsilon.$$

V práci [RJ] je diskutována problematika volby sumačních mezí x_1, x_2. Uvedené meze totiž závisí na hodnotě odhadovaného parametru p. Chceme-li najít nejmenší n takové, že odhad parametru p je dán se spolehlivostí $1-\alpha$, tj. zajistit, aby platilo

$$Pr\left(\left|\frac{X}{n} - p\right| < \varepsilon\right) = \sum_{x=x_1}^{x_2} \binom{n}{x} p^x (1-p)^{n-x} > 1 - \alpha,$$

obvykle se volí k určení x_1, x_2 odhad $p = \frac{1}{2}$. Jako důvod se uvádí maximální rozptyl o hodnotě $\frac{n}{4}$. Avšak uvedený argument není korektní. Například při volbě $n = 5$, p=0.4, ε=0.1 je jediným přirozeným číslem x_1 takovým, že $\frac{x_1}{n} = \frac{x_1}{5}$ v intervalu $(p-\varepsilon,\, p+\varepsilon) = \left(\frac{3}{2},\, \frac{5}{2}\right)$ číslo $x_1 = 2$. Tedy

$$Pr\left(\left|\frac{X}{n} - p\right| < \varepsilon\right) = \quad Pr\left(\left|\tfrac{X}{5} - 0.4\right| < 0.1\right) = \quad Pr\left(X = 2\right) =$$
$$\binom{5}{2} 0.4^2 (1-0.4)^3 \doteq \quad 0.3456.$$

Při doporučované volbě $p = 0.5$ ovšem získáme ($x_1 = 2$, $x_2 = 3$):

$$Pr\left(\left|\frac{X}{n} - p\right| < \varepsilon\right) = Pr\left(\left|\frac{X}{5} - 0.5\right| < 0.1\right) = Pr\left(X = 2\right) + Pr\left(X = 3\right) =$$
$$= \binom{5}{2} 0.5^2 (1-0.5)^3 + \binom{5}{3} 0.5^3 (1-0.5)^2 \doteq 0.625.$$

Maximální rozptyl tedy nezajišťuje vhodné stanovení spolehlivosti $1-\alpha$. Rovněž se ukazuje, že malá změna volby odhadu p může vést ke značné změně odpovídající pravděpodobnosti. V [RJ] autoři naznačují, že s uvedenou situací se lze vypořádat, pokud místo symetrického intervalového odhadu je použit asymetrický interval typu (10.1). Uvedený interval má odpovídat maximální pravděpodobnosti. Rozveďme úvahy autorů článku podrobněji.

Nechť p je odhadovaná pravděpodobnost a ε zadaná přesnost. Nejprve označme

$$D_p = \{\text{všechny intervaly } I \text{ takové, že } p \in I \text{ a } l\left(I\right) < 2\varepsilon\},$$

kde $l(I)$ označuje délku intervalu I. Pak minimální velikost vzorku n lze definovat jako takové nejmenší přirozené číslo n, pro které platí

$$\inf_p \left\{ \sup_{I \in D_p} \sum_{(k/n) \in I} Pr(X = k,\, p) \right\} > 1 - \alpha,$$

kde k je celé číslo a infimum je bráno přes uvažovaný rozsah hodnot odhadovaného parametru p binomické náhodné veličiny X.

Označme, pro reálné číslo x, symbolem $\lfloor x \rfloor$ největší celé číslo, které je menší nebo rovno než x.

Věta 5. *Mějme ε, α a n dané. Nechť $i = \lfloor 2n\varepsilon \rfloor$. Pro $j \in \{1, ..., n-i\}$, definujme*
$r_j = \frac{\binom{n}{j-1}}{\binom{n}{i+j}}$ *a* $p_j = \frac{r_j^{1/(i+1)}}{\left(1+r_j^{1/(i+1)}\right)}$.
Označme symbolem $H(p)$ hodnotu pravděpodobnosti intervalu s nejvyšší hustotou odpovídající p :

$$H(p) = \sup_{I \in D_p} \sum_{(k/n) \in I} Pr(X = k,\, p).$$

Pak

$$\inf \{H(p) : 0 \le p \le 1\} = \min (\{H(p_j) : 1 \le j \le n-i\}).$$

Věta 5 uvádí, že pro výpočet minimální oblasti přes $p \in [0, 1]$ s nejvyšší hustotou stačí zvažovat pouze $n - i$ hodnot p_j. Podobně pokud $p \le b$, pouze $H(p_j)$, kde $p_j < b$ musí být vypočítáno a délka vzorku je pak nejmenší n takové, že platí:

$$\min (\{H(p_j) : p_j < b\},\ H(b)) > 1 - \alpha.$$

Z věty 5 vychází následující algoritmus (viz str. 87, [RJ]):

1. Mějme dané ε, $b \le 0.5$ a α. Zvolíme počáteční odhad pro délku vzorku n. (Pokud $b > 0.5$, nahradíme b číslem $1 - b$.) Pro počáteční odhad n použijeme standardní vzorec (10.3) s volbou $p = b$.

2. Vypočítáme $i = \lfloor 2n\varepsilon \rfloor$,

$$r_j = \frac{\binom{n}{j-1}}{\binom{n}{i+j}}, \quad a \quad p_j = \frac{r_j^{1/(i+1)}}{\left(1 + r_j^{1/(i+1)}\right)}, \quad j = 1, 2, ..., n-i.$$

3. Vypočítejme $H(p_j) = \sum_{k=j}^{j+i} \binom{n}{k} p_j^k (1-p_j)^{n-k}$.

4. Označme $s = \max\{j : p_j \leq b\}$ a vypočítejme

$$H(b) = \sum_{k=s}^{s+i} \binom{n}{k} b^k (1-b)^{n-k}.$$

5. (a) Pokud neznáme předběžný odhad p, vypočítáme

$$p_{min} = \min\left(\{H(p_j) : \ 1 \leq j \leq n-i\}\right).$$

(b) Pokud $p < m < 0.5$, vypočítáme

$$p_{min} = \min\left(\{H(p_j) : \ p_j \leq b\}, H(b)\right).$$

6. Opakujeme kroky 2−5 s novými hodnotami pro n, dokud $p_{min} > 1-\alpha$ pro n, ale ne pro $n-1$.

Předchozí algoritmus lze snadno implementovat ve většině programovacích jazyků. V této práci byl zvolen Matlab.

10.2. **Program 3.** Následující funkce obsahuje jako první parametr požadovanou přesnost epsilon, druhým parametrem je hladina významnosti alpha a jako třetí parametr může být zadán předběžný horní odhad b odhadované pravděpodobnosti p. Není-li odhad předem znám, volíme $b = 1$, a tedy $0 < p(1-p) \leq 1/4$. Nejprve implementujme předběžný odhad n pomocí vztahu (10.3):

```
function[n]=klasickyOdhadN(epsilon,alpha,b)
  %epsilon predstavuje zadanou presnost
  %alpha je zadana hladina vyznamnosti
  %b je horni apriorni horni odhad parametru p
  Z=norminv(1-alpha/2);
  right=max(1/4,b*(1-b));
  n=ceil((Z^2*right)/epsilon^2);
end
```

Zavoláme-li uvedenou funkci s přesností $\varepsilon = 0.2$, hladinou významnosti $\alpha = 0.05$, získáváme v případě, že není znám předběžný odhad p (tzn. $b = 1$) následující hodnotu $n = 25$:

```
>> klasickyOdhadN(0.2,0.05,0.1)
ans =
    25
```

Nyní představme implementaci výše popsaného algoritmu:

```
function[pMin]=delkaVzorku(epsilon,b,n)
%funkce přijímající 3 vstupní hodnoty
% b - odhadované p, epsilon - šířka intervalu,
% n - délka vzorku, a 1 výstupní hodnotu,
% která nám udá, jakou hustotu pravděpodobnosti,
% vyplní vzorek se stejnými parametry, jako je funkce.
if b>0.5
b=1-b;
end
% zařizuje, ať odhadované p se nachází v intervalu <0,0.5>
i=floor(2*n*epsilon);
% floor = největší celé číslo, menší nebo rovno
i r=zeros(n-i,1);
p=zeros(n-i,1);
Hj=zeros(n-i,1);
%vytvořím si vektory s nulami
Hm=0;
s=0;
for j=1:n-i
r(j)=(nchoosek(n,j-1))/(nchoosek(n,i+j));
    p(j)=(r(j)^(1/(i+1)))/(1+r(j)^(1/(i+1)));
    for k=j:j+i
        Hj(j)=Hj(j)+(nchoosek(n,k)*
        (p(j)^k)*(1-p(j))^(n-k));
    end
%suma na výpočet Hj(j)
    if p(j)<=b
        s=j;
        %hledání maxima, splňující podmínku p(j)<=b
    end
end
%slouží k vypočítání r(j) a p(j)
if s==0
   disp('fail');
    %varovná zpráva, když žádná j nesplnila podmínku
end
```

```
for k=s:s+i
    Hm=Hm+(nchoosek(n,k)*b^k*(1-b)^(n-k));
end
% suma na výpočet Hm
if b==0.5
    pMin=min(Hj);
    % do pMin se zapíše minimum z vektoru Hj
else
    pMin=Hm;
%zapíšeme na pMin Hm
    for j = 1:n-i
        if pMin>Hj(j) && p(j)<=b
            pMin=Hj(j);
% hledáme minimum z vektoru Hj, splňující podmínky
end
end
end
% dvě možnosti, kudy se dále dáti
% a) pokud m je 0.5 b) pokud není
end
```

Využijeme-li předchozího předběžného odhadu $n = 25$, získáme aplikací výše uvedeného programu se vstupy $\varepsilon = 0.2$, $b = 0.5$, hodnotu $p_{min} = 0.9710$.

```
>> delkaVzorku(0.2,0.5,25)
ans =
    0.9710
```

Lze vidět, že p_{min} je relativně velká hodnota pro hladinu významnosti $\alpha = 0.05$. Aplikujeme znova výše uvedený program se stejnými vstupy kromě n, které zvolíme $n = 20$. Získáváme následující hodnotu: $p_{min} = 0.9534$. Horní odhad nejmenšího počtu pokusů nutných ke stanovení odhadu p se spolehlivostí 95% je tedy 20. Bohužel z práce [RJ] není možné jednoduše vyčíst, jak konkrétně stanovit příslušné asymetrické intervalové odhady p. Z uvedené práce však pravděpodobně vychází článek [XC], ve kterém jsou mimo jiné provedeny do důsledku symetrické intervalové odhady.

10.3. **Exaktní symetrické odhady pro binomické rozdělení.** V této části, založené především na výsledcích práce [RJ], představíme optimální

algoritmus k určení symetrického intervalového odhadu parametru p binomické náhodné veličiny s danou spolehlivostí a nejmenším možným n. Nejprve připomeňme a zaveďme některá nová značení:

Označme symbolem $\lceil x \rceil$ nejmenší celé číslo, které je větší nebo rovno x, obdobně označme symbolem $\lfloor x \rfloor$ největší celé číslo, které je menší nebo rovno x. Pro nezáporná celá čísla m kombinatorická funkce $\binom{m}{z}$ s ohledem na celé číslo z znamená:

$$\binom{m}{z} = \begin{cases} \frac{m!}{z!(m-z)!}, & 0 \leq z \leq m, \\ 0, & z < 0 \vee z > m. \end{cases}$$

Ještě si zaveďme binomickou funkci:

$$B(n,k,p) = \begin{cases} \binom{n}{k} p^k (1-p)^{n-k}, & 0 \leq k \leq n, \\ 0, & k < 0 \vee k > n. \end{cases}$$

Pro označení součtu binomických funkcí si zavedeme označení S.

$$S(n,k,l,p) = \sum_{i=k}^{l} B(n,i,p)$$

Věta 6. *Mějme $0 < \varepsilon < 1$ a $0 \leq a < b \leq 1$. Nechť $X_1, ..., X_n$ jsou totožné a nezávislé Bernoulliho náhodné proměnné takové, že pro $i = 1, ..., n$, $Pr\{X_i = 1\} = 1 - Pr\{X_i = 0\} = p$, u kterého platí $p \in [a,b]$. Nechť $p_n = \frac{\sum_{i=1}^{n} X_i}{n}$. Pak minimum $Pr\{|p_n - p| < \varepsilon\}$ s ohledem na $p \in [a,b]$ je dosaženo v konečné množině $\{a,b\} \cup \left\{\frac{\ell}{n} + \varepsilon \in (a,b) : \ell \in \mathbb{Z}\right\} \cup \left\{\frac{\ell}{n} - \varepsilon \in (a,b) : \ell \in \mathbb{Z}\right\}$, která má méně než $2n(b-a) + 4$ prvků.*

Skutečností symetrie, že $Pr\{|(1-p_n) - (1-p)| < \varepsilon\} = Pr\{|p_n - p| < \varepsilon\}$, můžeme omezit p na menší rozsah $[a', b']$ takové, že:

$$a' = \begin{cases} a & pro\ a + b \leq 1, \\ 1 - b & pro\ a + b > 1 \end{cases} \qquad b' = \begin{cases} b & pro\ b \leq \frac{1}{2}, \\ \frac{1}{2} & pro\ a < \frac{1}{2} < b, \\ 1 - a & pro\ a \geq \frac{1}{2}. \end{cases}$$

Je zřejmé, že $0 \leq a' < b' \leq \frac{1}{2}$ a $b' - a' \leq b - a$. Proto bez ztráty jakékoliv obecnosti můžeme předpokládat $0 \leq a < p < b \leq \frac{1}{2}$.

Věta 7. *Mějme* $0 \leq a < b \leq \frac{1}{2}$ *a* $0 < \varepsilon < \frac{1}{2}$. *Definujeme:*

$$\mathscr{S} = \{c_+(\ell)\,|1 + \lfloor n(a-\varepsilon)\rfloor \leq l \leq \lceil n(b-\varepsilon)\rceil - 1\} \cup \{c_a, c_b\}$$
$$\cup \{c_-(\ell)\,|1 + \lfloor n(a+\varepsilon)\rfloor \leq \ell \leq \lceil n(b+\varepsilon)\rceil - 1\},$$

kde

$$c_a = \sum_{k=\lfloor n(a-\varepsilon)\rfloor+1}^{\lceil n(a+\varepsilon)\rceil-1} B(n,k,a), \qquad c_b = \sum_{k=\lfloor n(b-\varepsilon)\rfloor+1}^{\lceil n(b+\varepsilon)\rceil-1} B(n,k,b),$$

$$c_+(\ell) = \sum_{k=\ell+1}^{\ell-1+\lceil 2n\varepsilon\rceil} B\left(n,k,\frac{\ell}{n}+\varepsilon\right), \qquad c_-(\ell) = \sum_{k=\ell+1-\lceil 2n\varepsilon\rceil}^{\ell-1} B\left(n,k,\frac{\ell}{n}-\varepsilon\right)$$

pro $\ell \in \mathbb{Z}$.

Pak následující tvrzení je/jsou pravdivé:

(I)*Minimum* $Pr\{|p_n - p| < \varepsilon\}$ *s ohledem na* $p \in [a,b]$ *je rovno minimu* $\mathscr{S}$ *tj.* $\min_{p\in[a,b]} Pr\{|p_n - p| < \varepsilon\} = \min \mathscr{S}$.

10.4. **Program 4.** V této práci byl implementován algoritmus inspirovaný větou 7 a vytvořen program ze dvou funkcí, které budou následovat v programu Matlab.

Následující funkce je funkce vedlejší, která vypočte binomickou funkci B s danými parametry n, k, p.

```
function[B]=spoctiB(n,k,p)
%tato funkce má 3 vstupní a 1 výstupní parametr
%n počet pokusů, k poč. úsp., p pravděpodobnost úsp.
%B výsledek binomické funkce
if k <0 || k>n
    B=0;
else
    %pokud k bude záporné nebo větší, než n,
    %tak považujeme 0 jako výsledek.
    B=nchoosek(n,k)*p^k*(1-p)^(n-k);
    %vypočítáme binomickou funkci
end
end
```

Následující funkce je funkce hlavní, která vypočte pravděpodobnost Pr, že se budeme nacházet v intervalu $[a, b]$ s přesností ε při délce vzorku n.

```
function[pr]=programX(a,b,epsilon,n)
%tato funkce má 4 vstupní a 1 výstupní parametr
%a - dolní odhad, b - horní odhad parametru p
%epsilon - šířka intervalu, n - počet pokusů
spodekCa=floor(n*(a-epsilon))+1;
vrsekCa=ceil(n*(a+epsilon))-1;
%spodni a horni mez sumy pro vypocet ca;
ca=0;
for k=spodekCa:vrsekCa
    ca=ca+spoctiB(n,k,a);
end
%vypocet ca;
spodekCb=floor(n*(b-epsilon))+1;
vrsekCb=ceil(n*(b+epsilon))-1;
%spodni a horni mez sumy pro vypocet cb;
cb=0;
for k=spodekCb:vrsekCb
    cb=cb+spoctiB(n,k,b);
end
%vypocet cb;
spodekCPlusL=1+floor(n*(a-epsilon));
vrsekCPlusL=ceil(n*(b-epsilon))-1;
%meze L pro VCPlusL;
it=1;
delkaVektoruCPlusL=vrsekCPlusL-spodekCPlusL+1;
VCPlusL=zeros(delkaVektoruCPlusL,1);
%výpočet délky vektoru a jeho vytvoření
for i=spodekCPlusL:vrsekCPlusL
    VCPlusL(it)=i;
    it=it+1;
end
%přiřazování čísla L do vektoru
spodekCMinusL=1+floor(n*(a+epsilon));
vrsekCMinusL=ceil(n*(b+epsilon))-1;
it=1;
```

```
delkaVektoruCMinusL=vrsekCMinusL-spodekCMinusL+1;
VCMinusL=zeros(delkaVektoruCMinusL,1);
for i=spodekCMinusL:vrsekCMinusL
    VCMinusL(it)=i;
    it=it+1;
end
%analogicky provedeme k výpočtu L pro VCMinusL;
cPlusL=zeros(delkaVektoruCPlusL,1);
for i=1:delkaVektoruCPlusL
    spodekForCPlusL=VCPlusL(i)+1;
    vrsekForCPlusL=VCPlusL(i)-1+ceil(2*n*epsilon);
    %meze pro výpočet cPlusL;
    for k=spodekForCPlusL:vrsekForCPlusL
       cPlusL(i)=cPlusL(i)+spoctiB(n,k,VCPlusL(i)/n+epsilon);
    end
    %výpočet cPlusL
end
cMinusL=zeros(delkaVektoruCMinusL,1);
for i=1:delkaVektoruCMinusL
    spodekForCMinusL=VCMinusL(i)+1-ceil(2*n*epsilon);
    vrsekForCMinusL=VCMinusL(i)-1;
    for k=spodekForCMinusL:vrsekForCMinusL
        cMinusL(i)=cMinusL(i)
        +spoctiB(n,k,VCMinusL(i)/n-epsilon);
    end
end
%analogicky pro výpočet cMinusL;
prvniPulS=union(cPlusL,cMinusL);
%sjednoceni vektoru cPlusL a cMinusL;
druhaPulS=union(ca,cb);
%sjednocení ca a cb;
S=union(prvniPulS,druhaPulS);
%do S uložíme sjednocení všech množin;
pr=min(S);
%pr zvolíme jako minimum množiny S;
end
```

Název funkce	α	ε	b	n	α	ε	b	n	α	ε	b	n
odhadN	0.05	0.05	0.5	385	0.05	0.075	0.5	171	0.05	0.1	0.5	97
delkaVzorku	0.05	0.05	0.5	370	0.05	0.075	0.5	160	0.05	0.1	0.5	90
programX	0.05	0.05	0.5	391	0.05	0.075	0.5	174	0.05	0.1	0.5	101
odhadN	0.05	0.05	0.3	323	0.05	0.15	0.49	43	0.05	0.1	0.35	88
delkaVzorku	0.05	0.05	0.3	310	0.05	0.15	0.49	37	0.05	0.1	0.35	80
programX	0.05	0.05	0.3	331	0.05	0.15	0.49	47	0.05	0.1	0.35	91
odhadN	0.05	0.05	0.1	139	0.1	0.1	0.3	57	0.4	0.02	0.5	443
delkaVzorku	0.05	0.05	0.1	120	0.1	0.1	0.3	50	0.4	0.02	0.5	425
programX	0.05	0.05	0.1	141	0.1	0.1	0.3	60	0.4	0.02	0.5	451

Tabulka 3. Porovnávání

Podle výsledků ve výše uvedené tabulce se potvrzuje, že s rostoucím počtem pokusů se odhady n pomocí aproximace normálním rozdělením (odhadN) příliš neliší od exaktních výsledků založených na algoritmu programX. Z tabulky je taky patrný vliv požadované přesnosti a spolehlivosti na odhad n.

11. Závěr

V první části práce byly připomenuty některé elementární poznatky týkající se aproximace binomického rozdělení, a to zejména rozdělením normálním. Méně známé výsledky jsou obsaženy ve druhé části práce. Ukazuje se, že k určení nejmenšího počtu Bernoulliho pokusů nutných k odhadu parametru p pravděpodobnosti úspěchu v jednom pokusu je algoritmus navržený v práci [XC] daleko vhodnější než standardní odhad vycházející z aproximace normálního rozdělení. Při dnešním stavu výpočetní techniky se výhody aproximačních odhadů (složitost algoritmu, doba výpočtu) ztrácejí, a naopak vyniká přesnost v určení příslušného nejmenšího počtu pokusů. V řadě případů tak lze podstatně zlevnit celý testovací proces.

Reference

[GD] *https://upload.wikimedia.org/wikipedia/commons/f/f4/Galton_board.png*

[GC] *https://upload.wikimedia.org/wikipedia/commons/thumb/c/c8/Gaussian_distribution.svg/2000px-Gaussian_distribution.svg.png*

[JB] https://upload.wikimedia.org/wikipedia/commons/b/bb/Jakob_Bernoulli.jpeg

[AM] *https://upload.wikimedia.org/wikipedia/commons/1/1b/Abraham_de_moivre.jpg*

[FG] *https://upload.wikimedia.org/wikipedia/commons/0/0b/Francis_Galton.jpg*

[WQ] *https://en.wikiquote.org/wiki/Main_Page*

[WEN] *https://en.wikipedia.org/wiki/Main_Page*

[WCS] *https://cs.wikipedia.org/wiki/Hlavn%C3%AD_strana*

[ML] Litschmannová, Martina. Vybrané kapitoly z pravděpodobnosti. Ostrava, 2011. Dostupné z *http://mi21.vsb.cz/sites/mi21.vsb.cz/files/unit/vybrane_kapitoly_pravdepodobnost.pdf*

[ZR] Riečanová, Zdena. Numerické metódy a matematická štatistika, Bratislava, Alfa, 1987.

[PR] *ftp://math.feld.cvut.cz/pub/prucha/m3c/predn/pravd/u4.pdf*

[XC] *https://arxiv.org/pdf/0707.2113.pdf*

[RJ] http://www.medicine.mcgill.ca/epidemiology/Joseph/publications/Methodological/binexact.pdf

[GS] Grinstead ch. M.,Snell L.: Introduction to Probability, AMS, 2nd edition, 2003

[AJ] Anděl J.: Matematická statistika, SNTL, Praha, 1985